University of Science and Technology

College of Graduate Studies & Academic Advancement

A Comparison Study of Machine Learning Algorithms for Phishing Email Detection

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Masters of Computer Science

Student Name: Mohammed Abd-Elmonem Ahmed Abd-Elrahim

Supervisor Name: Dr. Altayeb Altaher

2012

Abstract

There are many applications available for phishing detection. However, unlike predicting spam, there are only few studies that compare machine learning techniques in predicting phishing. The present research compares the predictive accuracy in many measurements including true positive rate (TPR), false positive rate (FPR), Precision, and F-Measure for the several machine learning methods including Multilayer Perceptron (MLP), Support Vector Machines (SVM), Random Tree (RT), Random Forest (FR), Logistic Regression (LR), Naïve Bayesian (NB), AdaBoost (AB), and J48 for predicting phishing emails. A data set of 8266 phishing and legitimate emails was used in the comparative study. In addition, three sets of the features were used to train and test the classifiers, these sets are: 47 features, binary features, and 20 features selected by the information gain. The results show that the RF algorithm outperforms the all others classifiers in true positive, false positive, precision, F-measure, While AB algorithm had worst rates in all measurements. Although the set of 20 features selected by the information gain present the better results on the Multilayer Perceptron (MLP), Random Tree (RT), Logistic Regression (LR), and J48 algorithms, while Random Forest (RF), Support Vector Machine (SMO), and Naïve Bayesian (NB) were presents a good results in 47 features, rather than all algorithms has worst results on the set binary features set.

المستخلص

هنالك العديد من التطبيقات المستخدمة في عملية إكتشاف التصيد عن طريق رسائل البريد. علي كُلاً، الدراسات التي تقارن بين الخوارزميات التي تعرف بـ(Machine Learning) والتي تستخدم في إكتشاف التصيد عن طريق البريد قليل جدا مقارنة بالتي أُجريت لإكتشاف البريد العشوائي (spam). هذا البحث يقارن دقة التنبؤ او الإكتشاف للتصيد لمجموعة من الخوارزميات والخوارزميات هي Multilayer Perceptron (MLP), Support Vector Machines (SVM), Random Tree (RT), Random Forest (FR), Logistic Regression (LR), Naïve Bayesian (NB), AdaBoost (AB), and J48.

و هذه المقارنات تمت في عُدة قياسات وهي: true positive rate (TPR), false positive rate (FPR), Precision, F-Measure. تم إستخدام ملفات لرسائل تصيد واخري معتادة بلغ عددها 8266 لكي تختبر عليها الخوارزميات المُراد مقارنتها، بالإضافة لآنه تم إجراء ثلاثة تجارب كل تجربة تأخذ مجموعة مختلفة من الخصائص الخاصة برسائل البريد، حيث تم إختيار جميع الخصائص والبالغ عددها 47 في التجربة الاولي، وكل الخصائص التي تحمل قيم ثنائية إما 0 او 1 في التجربة الثانية، بينما تم إختيار 20 من المجموع الكُلي للخصائص في التجربة الثالثة، هذه الخصائص هي التي تملك اعلي نسبة تأثير علي عملية التنبؤ، وقد تم إختيار هذه الخصائص عن طريق إستخدام ما يعرف بـ (the Information Gain).

ومن النتائج وجد أن خوارزمية RF هي الافضل علي الإطلاق من بين الخوارزميات بعدما كانت الافضل في كل القياسات، بينما كانت خوارزمية AB هي الاسواء في كل القياسات. وايضاً كانت نتائج الخوارزميات Multilayer Perceptron (MLP), Random Tree (RT), Logistic Regression

(LR), and J48 في التجربة الثالثة (الخصائص المختارة ذات التأثير العالي) هي الافضل مقارنتاً بنتائجها في التجاربتين الأخرتين.

فيما كانت الخوارزميات Random Forest (RF), Support Vector Machine (SMO), and Naïve Bayesian (NB) ذات نتائجها افضل في التحربة الاولي (كل الخصائص). وكانت كل الخوارزميات قد اعطت اسواء نتايجها في التجربة الثانية(الخصائص ذات القيم الثنائية).

1: Introduction

This chapter presents an overview of email and phishing, research problem definition, objective of research, research questions, research methodology, and thesis structure.

Electronic Mail

Electronic mail, widely known as email, is an electronic equivalent of postal mail. Email is mainly used to compose, send, forward, receive, and archive the messages over the electronic communication systems. Email started in 1965 as a way of communication between multiple users on time-sharing main frame computers. Though it is difficult to identify the exact origin of email, early systems are known to provide such facility. In 1971 Ray Tomlinson first used '@' symbol for the email address to separate the names of the user and their machines. He wrote the first email system for users on different hosts connected to the ARPANET [17]. Thereafter, the emergence of global Internet significantly increased the popularity of the email as a result, and it evolved into a pervasive application. The current email system is based on the SMTP protocol RFC 821 and 822 developed in 1982 and extended in RFC 2821 in 2001 [13]. This system defines a common standard to unite the different messaging protocols in existence prior to 1982. It allowed users the ability to exchange messages with one another using a system based on the SMTP protocol and email addresses. These protocols allowed messages to flow from one user to another, making it practical and easy for different users to communicate independent of the service-provider or the client application.

Email Phishing

Phishing is a form of identity theft that occurs when a malicious Web site impersonates a legitimate one in order to acquire sensitive information such as passwords, account details, or credit card numbers. Phishing is a deception technique that utilizes a combination of social engineering and technology to gather sensitive and personal information, such as passwords and credit card details by masquerading as a trustworthy person or business in an electronic communication. Phishing makes use of spoofed emails that are made to look authentic and purported to be coming from legitimate sources like financial institutions, ecommerce sites etc., to lure users to visit fraudulent websites through links provided in the phishing email.

The fraudulent websites are designed to mimic the look of a real company webpage. The phishing attacker's trick users by employing different social engineering tactics such as threatening to suspend user accounts if they do not complete the account update process, provide other information to validate their accounts or some other reasons to get the users to visit their spoofed web pages.

A Quick History of Phishing

Phishing is far from a new phenomenon. About a decade ago the first phishing incidents involving AOL (**A**merican **O**n**l**ine mail) account theft were reported. A few years later fake Hotmail login sites were a popular method for identity theft. Fortunately, the effects of these attacks on the victim were above all of an annoying nature and mostly did not yield substantial financial consequences. Meanwhile, as the internet evolved, the services offered became more mature.

A few years ago financial institutions started offering internet banking and online payment systems. By now, these systems have been widely accepted and it is estimated that about 70% of

the Dutch internet users make use of internet banking systems [18]. Consequently, attacks on these systems became interesting for organized crime.

In contrast to early phishing attacks, which were mostly only annoying to the victim, the phishing phenomenon had turned into a serious felonious business mainly targeting financial services. According to the Anti-Phishing Working Group [], there were 18,480 unique phishing attacks and 9666 unique phishing sites reported in March 2006. Phishing attacks affect millions of internet users and are a huge cost burden for businesses and victims of phishing. Gartner research conducted in April 2004 found that information given to spoofed websites resulted in direct losses for U.S. banks and credit card issuers to the amount of $1.2 billion [9]. Phishing has become a significant threat to users and businesses alike. Among 1.3 million complaints received by the Federal Trade Commission in 2009, identity theft ranked first and accounted for 21% of the complaints costing consumers over $1.7 [19]. Over the past few years, much attention has been paid to the issue of security and privacy. Existing literature dealing with the problem of phishing is scarce.

Problem Definition

Today there are many different techniques to prevent from the phishing Email; some of these techniques used the algorithms of machine learning. However, determining which of these algorithms is better is difficult because each of these algorithms may be optimal for some purpose.

Research Objective

The phishing Email detection is become one of the issues which supported by the company that provide Email service, the peoples use this service may thinking about different issues like accuracy and performance and so on. The objective of this research is to study the machine learning algorithms that are used in phishing detection, and indicate which of these Algorithms is optimal in which measurement, and what the better features are used to detect the phishing email.

Research Questions

The problem description leads to the principal research question of this study:

- What is the optimal technique to detect phishing Email?

When this question is asked, there is question suppose to be asked. Answer this question firstly to explain some terms, after that can answer the main question.

- What are the measurements used to evaluate the algorithms using in phishing detection techniques?
- What is set of used features that effectiveness in the phishing email detection?

Research Methodology

The proposed methodology is to take three sets of the selected features, and construct three experiments on each set using the eight commonly used algorithms in phishing Email detection and compare the result of the three experiments in different measurements to find which of this algorithm is better in which measurement, and .

Thesis Structure

This thesis introduces the related works on the area of this research in Chapter 2, introduces the used machine learning algorithms in Chapter 3, Chapter 4 presents the thesis methodology, features selected to experiment, and discussion of the measurements that is compared depends on, Chapter 5 presents and explains the results of the experiments, and concludes the research results and recommendation for future studies in Chapter 6.